

## **Лабораторная работа №5**

### **Трансформация данных.**

**Цель работы:** изучить процесс трансформации данных.

**Задачи работы:**

- изучить методику трансформации данных в Deductor Studio;
- изучить приведенные в лабораторной работе примеры;
- выполнить контрольное задание.

#### **1. Краткая теория**

##### **1.1. Трансформация данных**

Один из этапов подготовки данных к анализу – трансформация данных.

Каждая выборка исходных данных характеризуется набором свойств, которые могут повлиять на эффективность работы модели и снизить достоверность результатов анализа. Даже если данные очищены от таких факторов, ухудшающих их качество, как дубликаты, противоречия, шумы, аномальные значения, пропуски и др., они все еще могут не соответствовать методике и целям анализа. Это связано не с содержанием данных, а с их представлением и внутренней организацией. Данные могут быть разобщены, неупорядочены, представлены в форматах, с которыми не работает тот или иной алгоритм. Трансформация данных, то есть их преобразование к определенному представлению, формату или виду, оптимальному с точки зрения решаемой задачи, и призвана решить эту проблему.

Трансформация — это процесс оптимизации их представления и организации с точки зрения определенного метода анализа. Трансформация данных зависит от задач, алгоритмов и целей анализа. То есть для одной задачи могут потребоваться одни виды трансформации, а для другой — другие.

Трансформация данных — комплекс методов и алгоритмов, направленных на оптимизацию представления и форматов данных с точки зрения решаемых задач и целей анализа. Трансформация данных не ставит целью изменить информационное содержание данных. Ее задача — представить эту информацию в таком виде, чтобы она могла быть использована наиболее эффективно.

Основными целями трансформации данных на этапе процесса ETL являются приведение их в соответствие с моделью данных, используемой в хранилище, осуществление корректной консолидации и собственно загрузка в хранилище.

##### **1.2. Основные методы трансформации данных.**

**Преобразование временных данных.** Позволяет оптимизировать представление таких данных с целью обеспечения дальнейшего анализа, например решения задачи прогнозирования временного ряда или группировки по временному периоду.

**Квантование.** Позволяет разбить диапазон возможных значений числового признака на заданное количество интервалов и присвоить номера интервалов или иные метки попавшим в них значениям.

**Сортировка.** Позволяет изменить порядок следования записей исходной выборки данных в соответствии с алгоритмом, определенным пользователем. В некоторых случаях сортировка дает возможность упростить визуальный анализ выборки, оперативно определить наибольшие и наименьшие значения признаков и т.д.

**Слияние.** Позволяет объединить две таблицы по одноименным полям или дополнить одну таблицу записями из другой, которые отсутствуют в дополняемой. Слияние применяется в тех случаях, когда информацию в анализируемой выборке данных необходимо дополнить информацией из другой выборки. При объединении к записям исходной выборки добавляются все записи другой. В случае дополнения к исходной выборке добавляются только те данные, которые отсутствовали в исходной. Операция слияния является одним из способов обогащения данных: если выборка содержит недостаточно данных для анализа, то ее можно дополнить недостающей информацией из другой выборки.

**Группировка.** Очень часто информация, интересующая аналитика, в таблице оказывается «разбавлена» посторонними данными, разобщена, разбросана по отдельным полям и записям. Используя группировку, можно обобщить нужную информацию, объединить ее в минимально необходимое количество полей и значений. Обычно предусматривают возможность выполнения и обратной операции — разгруппировки.

**Настройка набора данных.** Позволяет изменять имена, типы, метки и назначения полей исходной выборки данных. Например, если поле, содержащее числовую информацию, в источнике данных по какой-либо причине имеет строковый тип, значения этого поля не могут обрабатываться как числа. Чтобы работа с числовыми данными этого поля стала возможной, их следует преобразовать к числовому типу.

**Табличная подстановка значений.** Позволяет производить замену значений в исходной выборке данных на основе так называемой таблицы подстановки. Таблица подстановки содержит пары «исходное значение — новое значение». Каждое значение выборки данных проверяется на соответствие исходному значению таблицы подстановки, и если такое соответствие найдено, то значение выборки изменяется на соответствующее новое значение из таблицы подстановки. Это очень удобный способ для автоматической корректировки значений.

**Вычисляемые значения.** Иногда для анализа требуется информация, которая отсутствует в явном виде в исходных данных, но может быть получена на основе вычислений над имеющимися значениями. Например, если известны цена и количество товара, то сумма может быть рассчитана как их произведение. Для этих целей в аналитическое приложение включается своего рода калькулятор, который позволяет выполнять над

данными исходной выборки различные вычисления. Поскольку анализируемые данные могут быть различных типов (строковый, числовой, дата/время, логический), то механизм расчетов должен поддерживать работу не только с числовыми данными, но и с данными других типов, например выделять подстроку, выполнять логические операции и т. д.

**Нормализация.** Нормализация позволяет преобразовать диапазон изменения значений числового признака в другой диапазон, более удобный для применения к данным тех или иных аналитических алгоритмов, а также согласовать диапазоны изменений различных признаков. Часто используется приведение к единице, когда весь имеющийся диапазон данных «сжимается» в интервал  $[0; 1]$  или  $[-1; 1]$ . Особенно важно произвести правильную нормализацию данных в алгоритмах Data Mining, которые основаны на измерении расстояния между векторами объектов в многомерном пространстве признаков (например, в кластеризации).

Целью трансформации временных рядов является не изменение их содержания, а представление информации таким образом, чтобы обеспечивалась максимальная эффективность решения определенной задачи анализа. Можно выделить два основных типа преобразования, которые наиболее часто используются при подготовке временных рядов к анализу.

Скользящее окно применяется при решении задач прогнозирования и классификации состояний бизнес-объектов, чтобы преобразовывать последовательность значений ряда в таблицу, которую можно использовать для построения моделей или какой-либо другой обработки.

Преобразование даты и времени заключается в приведении даты и времени к виду, наиболее удобному для визуального анализа и обработки временного ряда. При этом результаты преобразования даты уже не являются значениями типа Дата/Время и могут обрабатываться как обычные числа и строки.

Скользящее окно применяется при обработке временных рядов, например, чтобы построить модель прогноза временного ряда. Целью прогнозирования значений временного ряда является предсказание значения  $x(n + 1)$  на основе предыдущих значений признака. Решение задачи прогнозирования возможно только в том случае, если значения временного ряда связаны между собой.

Для построения прогностической модели необходимо знать три параметра.

Интервал прогноза — временной интервал, на котором будет осуществляться прогнозирование: день, неделя, месяц, квартал или год.

Горизонт прогноза — на какое количество интервалов (дней, недель и т. д.) мы хотим получить прогноз.

Глубина погружения — количество значений интервалов прогноза в прошлом, которое мы будем использовать для предсказания значений интервалов в будущем.

При выборе интервала прогноза, возможно, понадобится агрегирование данных внутри интервала. Например, если в базе данных информация о продажах представлена по дням, то для построения прогноза по неделям придется объединить информацию за отдельные дни, вычислив сумму, среднее значение или используя другую функцию агрегации.

Если в качестве интервала прогноза выбрана неделя, то глубина погружения — количество недель в прошлом, значения продаж за которые мы будем использовать в качестве исходных данных для предсказания.

Горизонт прогноза в этом случае — количество недель, значения продаж за которые мы хотим предсказать.

При выборе глубины погружения и горизонта прогноза следует руководствоваться следующими соображениями.

Модель прогноза функционирует по принципу обобщения, то есть вывод о возможном значении в будущем делается на основе анализа большого числа значений из прошлого. Следовательно, глубина погружения должна в несколько раз превышать горизонт прогноза.

Чем большее число значений из прошлого используется для прогнозирования, тем выше степень обобщения и тем достовернее результаты предсказания. Однако при этом есть два ограничения. Во-первых, глубина погружения должна быть такой, чтобы значения из прошлого, используемые для прогнозирования, оставались актуальными, то есть правильно отражали текущее поведение исследуемого процесса.

Использование для построения модели прогноза слишком старых данных, утративших актуальность, приведет к снижению достоверности или к получению некорректных результатов. Во-вторых, слишком большая глубина погружения увеличит размерность выборки данных, полученной в результате обработки ряда скользящим окном, что повлечет за собой усложнение модели. Кроме того, возрастут временные и вычислительные затраты на анализ данных.

Чем дальше простирается горизонт прогноза, тем ниже достоверность результатов.

Группировка данных — полезный инструмент, применение которого в процессе подготовки данных к анализу позволяет более эффективно использовать содержащуюся в данных информацию.

Перед тем как приступить к группировке, необходимо определить в исходной выборке данных поля измерений и фактов. Измерения — это данные, характеризующие исследуемый процесс качественно, а факты — количественно. Например, процесс продаж описывается количественными и качественными данными. Качественные данные — наименования товаров и клиентов. Очевидно, что просто указать эти параметры недостаточно. С каждым проданным товаром или клиентом связан набор числовых показателей — фактов: цены, количества, суммы, скидки, наценки и т. д.

Каждое наименование товара или клиента, содержащееся в поле измерения, называется значением измерения. Суть группировки заключается

в том, что все записи, содержащие одноименные значения измерения, по которому производится группировка, объединяются в одну, а соответствующие факты агрегируются.

Рассмотрим выборку данных, отражающих продажу стройматериалов (рисунок 1). Она содержит измерения Дата, Клиент и Товар, а также факты Цена, Количество и Сумма.

Дата	Клиент	Товар	Цена	Количество	Сумма
01.03.2007	ООО «Полигон»	Цемент	150	20	3000
01.03.2007	ЗАО «Монтажник»	Керамзит	100	60	6000
01.03.2007	ООО «Тандем»	Кирпич	1500	5	7500
02.03.2007	ООО «Шплинт»	Плиты	1100	10	11 000
02.03.2007	ЗАО «Монтажник»	Блоки	900	20	18 000
02.03.2007	ООО «Полигон»	Кирпич	1500	10	15 000
03.03.2007	ООО «Агрострой»	Плиты	1100	8	8800
03.03.2007	ЗАО «Монтажник»	Керамзит	100	30	3000
03.03.2007	ООО «Тандем»	Блоки	900	10	9000
03.03.2007	ООО «Полигон»	Плиты	1100	30	33 000
04.03.2007	ООО «Шплинт»	Керамзит	100	100	10 000
04.03.2007	ООО «Тандем»	Кирпич	1500	10	15 000
04.03.2007	ЗАО «Монтажник»	Цемент	150	20	3000
04.03.2007	ООО «Полигон»	Блоки	900	15	13 500
05.03.2007	ООО «Агрострой»	Плиты	1100	20	22 000
05.03.2007	ООО «Шплинт»	Керамзит	100	50	5000
05.03.2007	ЗАО «Монтажник»	Цемент	150	40	6000
05.03.2007	ООО «Тандем»	Блоки	900	15	13 500
06.03.2007	ООО «Полигон»	Кирпич	1500	6	9000

Рисунок 1 – Выборка данных по продаже стройматериалов

Группировка может производиться по одному из трех измерений. Если в качестве измерения для группировки выбирается Дата, то все записи, содержащие одинаковые даты, будут объединены в одну (рисунок 2).

Дата	Цена	Количество	Сумма
01.03.2007	583,33	85	16 500
02.03.2007	1166,67	40	44 000
03.03.2007	800,00	78	53 800
04.03.2007	662,50	145	41 500
05.03.2007	562,50	125	46 500
06.03.2007	583,33	106	20 500

Рисунок 2 – Вариант группировки 1

При этом для измерения Цена была выбрана функция агрегации с вычислением среднего значения, для измерений Количество и Сумма — с вычислением суммы. Таким образом, этот вариант группировки позволяет получить информацию о том, какое количество единиц товара, по какой средней цене и на какую сумму было продано в пределах каждой даты.

Группировка по измерению Клиент позволяет получить те же агрегированные факты, что и в предыдущем примере, но уже не для каждой даты, а для каждого клиента (рисунок 3).

Клиент	Цена	Количество	Сумма
ЗАО «Монтажник»	258,33	200	40 500
ООО «Агрострой»	766,67	98	37 800
ООО «Полигон»	1030,00	81	73 500
ООО «Тандем»	1200,00	40	45 000
ООО «Шплинт»	433,33	160	26 000

Рисунок 3 – Вариант группировки 2

Группировка по измерению Товар позволяет получить информацию о том, по какой цене, в каком количестве и на какую сумму был продан каждый из товаров (рисунок 4).

Товар	Цена	Количество	Сумма
Блоки	900	60	54 000
Керамзит	100	310	31 000
Кирпич	1500	31	46 500
Плиты	1100	68	74 800
Цемент	150	110	16 500

Рисунок 4 – Вариант группировки 3

При группировке данных могут использоваться различные функции агрегации. К наиболее типичным относятся:

- Сумма – вычисляется сумма агрегируемых значений фактов.
- Среднее – вычисляется среднее агрегируемых значений фактов.
- Количество – это число агрегируемых значений фактов для каждой комбинации измерений.
- Максимум, минимум – максимальное или минимальное из агрегируемых значений.
- Медиана – агрегируемые значения сортируются в порядке возрастания, и из полученного набора выбирается центральное (т.е. такое значение, что все значения слева от него будут меньше, а справа – больше). Медиана – это порядковая статистика, использующаяся как альтернатива среднего значения, устойчивая к аномальным значениям данных. Если аномальное значение попадет в число усредняемых, то это может существенно сместить полученную оценку, в то время как медиана в большинстве случаев дает более устойчивую оценку.

Для строковых значений в качестве функций агрегации могут использоваться только максимум, минимум, количество, первый, последний. При этом максимум (минимум) нескольких строковых значений рассчитывается посимвольным сравнением. Сначала сравниваются два первых символа строк. Если их коды одинаковы, сравниваются вторые символы и т. д. Как только в строках появляется первый несовпадающий символ, функция агрегации принимает значение строки, код символа в которой оказался больше (меньше).

За счет объединения значений измерений группировка позволяет оптимизировать представление анализируемых данных с точки зрения эффективности анализа и интерпретируемости его результатов. Кроме того, группировка дает возможность снизить количество наблюдений, которые необходимо обработать в процессе анализа, а значит, уменьшить время и вычислительные затраты на его выполнение.

#### **Слияние данных.**

В практике анализа достаточно часто встречается ситуация, когда требуемые данные приходится собирать из нескольких таблиц. Необходимость в этом обычно возникает в следующих случаях.

- Данные, которые нужны для анализа, «разбросаны» по нескольким таблицам.
- Данные в исходной таблице несут недостаточно информации для качественного анализа, и поэтому требуется процедура их обогащения, которая обычно связана с добавлением в таблицу данных из сторонних источников.

Ситуация, когда анализируемые данные оказываются в нескольких таблицах или берутся из отдельных источников, а не из централизованного хранилища данных, может быть следствием непродуманного процесса консолидации и интегрирования данных на этапе ETL. Недостаточно

информативная выборка также встречается довольно часто. И дело здесь даже не столько в самой выборке, сколько в методике анализа, для которой она используется. Например, изначально множество данных предполагалось применять как обучающую выборку для построения модели прогноза, но впоследствии возникла необходимость в ее использовании для классификации объектов множества. Возможно, для решения задачи прогнозирования во множестве использовались один или два выходных признака, тогда как для надежной классификации требуется не менее трех или четырех. В этом случае недостающие признаки можно будет взять из другой таблицы. Таким образом, если для решения одной аналитической задачи информативность множества данных может быть достаточна или даже избыточна, то для другой она окажется недостаточной и потребует обогащения.

При необходимости объединить данные выполняется процедура слияния (merge). Таблица, к которой в процессе слияния добавляются данные из другой, называется исходной, или входящей; вторую таблицу, данные из которой добавляются к исходной, часто называют связываемой. Исходная и связываемая таблицы должны иметь одно или несколько одинаковых полей, на основе которых будет производиться связывание двух таблиц, — это поля связи.

Остальные поля, уникальные для каждой из таблиц, могут быть присоединены к результирующему набору после слияния.

Существует несколько способов слияния, которые применяются в зависимости от того, какие данные и в каком виде должны быть объединены в результирующей таблице.

Объединение применяется в тех случаях, когда к строкам исходной таблицы требуется добавить все строки связываемой, при этом строки добавляются снизу (рисунок 5).

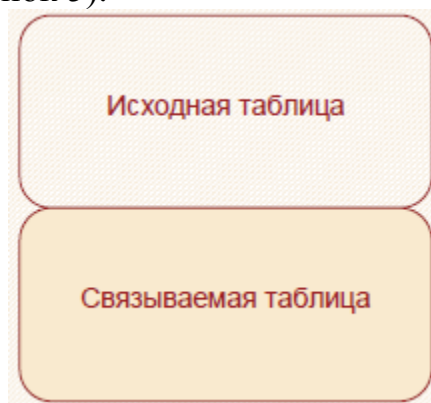


Рисунок 5 – Объединение таблиц

Внутреннее соединение позволяет получить в результирующем наборе только те записи, для которых значения в одном из полей связи совпадают. То есть в таблице, полученной в результате внутреннего соединения, останутся только те записи, которые содержат одинаковые значения в



заданном поле (или заданных полях). Принцип внутреннего соединения схематично представлен на рисунке 6.

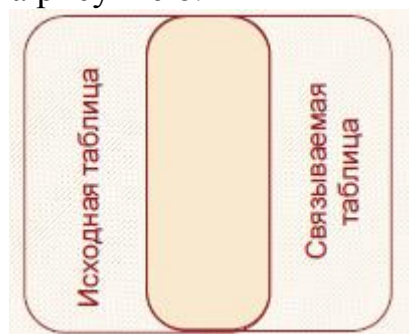


Рисунок 6 – Внутреннее соединение

При внешнем соединении все записи одной таблицы дополняются значениями из другой, если значения этих записей по ключевым полям совпадают. То есть таблицы связываются по полю Товар, и если существуют записи, где значения данного поля в обеих таблицах идентичны, то записи будут дополнены значениями, которые отсутствуют в одной таблице и присутствуют в другой. Фактически, этот механизм позволяет добавлять поля из одной таблицы в другую, но не по всем записям, а только по тем, значения которых в поле связи совпадают для обеих таблиц.

Кроме того, различают левое и правое внешнее соединение. При левом записи исходной таблицы дополняются значениями из связанной таблицы, а при правом — наоборот.

Схематично принцип внешнего соединения поясняется на рисунке 7.



Рисунок 7 – Внешнее соединение

Полное внешнее соединение. В результирующий набор включаются все строки и поля как исходной, так и связываемой таблиц. При этом, если в некоторой записи значения поля связи для обеих таблиц совпадают, то все поля этой записи заполняются соответствующими значениями. Если совпадение по полю связи отсутствует, то остальные поля такой записи будут заполнены пустыми значениями. Фактически это означает, что для значений исходной таблицы отсутствуют соответствующие значения в связываемой. Принцип работы полного внешнего соединения поясняется на рисунке 8.

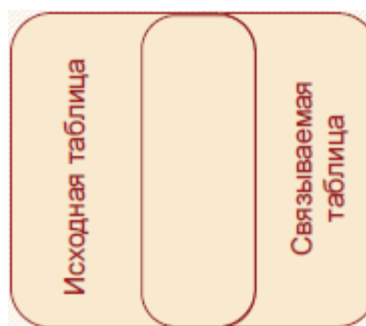


Рисунок 8 – Полное внешнее соединение

Одним из преобразований, которые часто используются при подготовке данных к анализу, является квантование. В основе операции квантования лежит процедура, состоящая из двух шагов.

- Диапазон значений, в пределах которого изменяется некоторая числовая величина (признак, показатель и т. д.), разбивается на заданное количество интервалов, каждому из которых присваивается определенный номер. Эти интервалы называются интервалами квантования, а присвоенные им номера — уровнями квантования.

- Каждое значение заменяется номером интервала квантования, в который попало данное значение.

Графически процесс квантования может быть представлен в следующем виде (рисунок 9).

На рисунке штрихпунктирными линиями представлены границы интервалов квантования, справа расположены номера интервалов, а внизу — результирующие значения, которые будут присвоены наблюдениям в результате квантования.

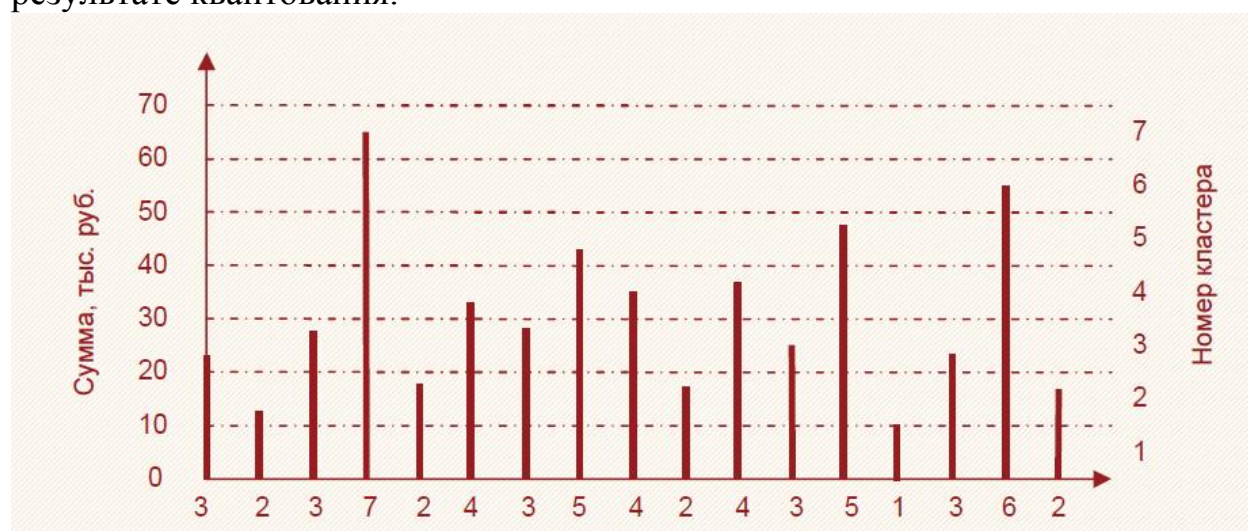


Рисунок 9 – Процесс квантования

Квантование широко используется во всех областях, где возникает необходимость в обработке, передаче и хранении данных. Квантование —

неотъемлемая часть процесса преобразования аналоговых (то есть непрерывных по времени и амплитуде) сигналов в цифровые (то есть дискретные по времени и квантованные по амплитуде). Квантование позволяет представлять и хранить данные в более компактном и защищенном от искажений виде. Процесс дискретизации заключается в представлении непрерывной функции в виде набора отдельных значений, взятых в определенные моменты времени, — отсчеты. В результате квантования значения отсчетов преобразуются в номера интервалов квантования, в которые эти значения попали.

Принцип преобразования аналоговых данных в цифровые представлен на рисунке 10.

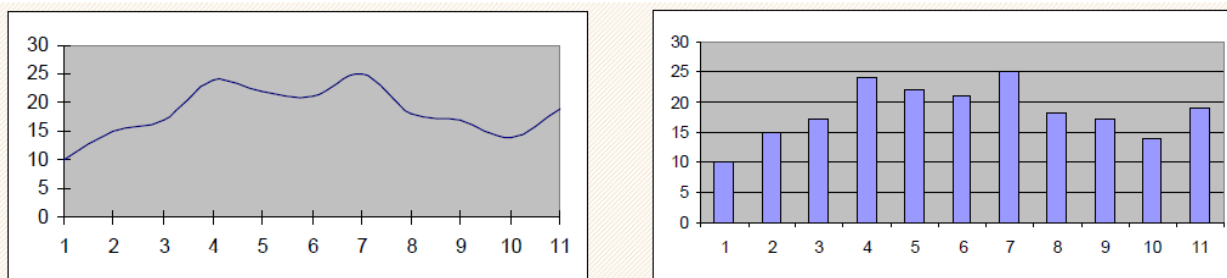


Рисунок 10 – Иллюстрация принципа преобразования

Выбрать количество интервалов квантования можно исходя из следующих соображений.

- Если квантование выполняется для преобразования непрерывных данных в дискретные, то число интервалов будет определяться числом уникальных значений (меток, категорий), которое используется при решении задачи анализа.

- Необходимо учитывать требуемую точность описания данных. Например, может быть поставлено условие, что количество интервалов квантования должно быть таким, чтобы ширина интервала не превышала 10 % от полного диапазона изменения исходных значений.

Иногда может потребоваться проведение экспериментов, чтобы определить лучшие параметры квантования с точки зрения решения конкретной задачи анализа.

Различают два основных метода квантования:

- равномерное (однородное) квантование;
- неравномерное (неоднородное) квантование.

При равномерном квантовании диапазон изменения значений признака разделяется на интервалы одинаковой ширины, а при неравномерном ширина интервалов может быть различной.

Первый метод используется, если данные равномерно распределены по всему диапазону их изменения, то есть в результате квантования не будет интервалов, в которых значения почти отсутствуют или заполнены очень плотно. В противном случае лучшие результаты даст второй метод.

Транспонирование – это термин из теории матриц, который обозначает операцию, преобразующую столбцы матрицы в строки, а строки – в столбцы. При работе с таблицами, содержащими анализируемые данные, этот термин имеет более широкий смысл.

Транспонирование на этапе трансформации используется для оптимизации структуры источника данных с точки зрения определенной задачи. С помощью транспонирования можно не только менять местами строки и столбцы таблицы, но и производить более сложные манипуляции с ее структурой.

### **Пример №1 «Группировка и Слияние с узлом».**

В данном примере используются наборы данных sales.ddf и plan.ddf (находятся в папке «К лабораторной работе №5»).

В файле sales.ddf находятся продажи строительных товаров с указанием из названия, товарной группы, количества, суммы, единицы измерения, а также информации, в каком городе произведена продажа и какому типу клиента. Всего 98471 запись.









№	Поле	Значение
1	 Дата	01.03.2004
2	 Количество	28
3	 Сумма с учетом скидки	13708,8
4	 Группа товара	Напольные покрытия
5	 Группа клиента	Клиент
6	 Единица измерения	шт
7	 Город	Балашиха
8	 Товар	Доска паркетная UPOFLOOR Лос 2085x188x14 дуб

Рисунок 11 – Содержимое файла sales.ddf

Поставим задачу сделать набор данных с продажами по дням по всем городам. Поскольку в один и тот же день конкретный товар мог быть продан несколько раз, необходимо сделать группировку.

Нужен узел Группировка. Он позволяет объединять записи по полям-измерениям, агрегируя данные в полях-фактах для дальнейшего анализа (рисунок 12).

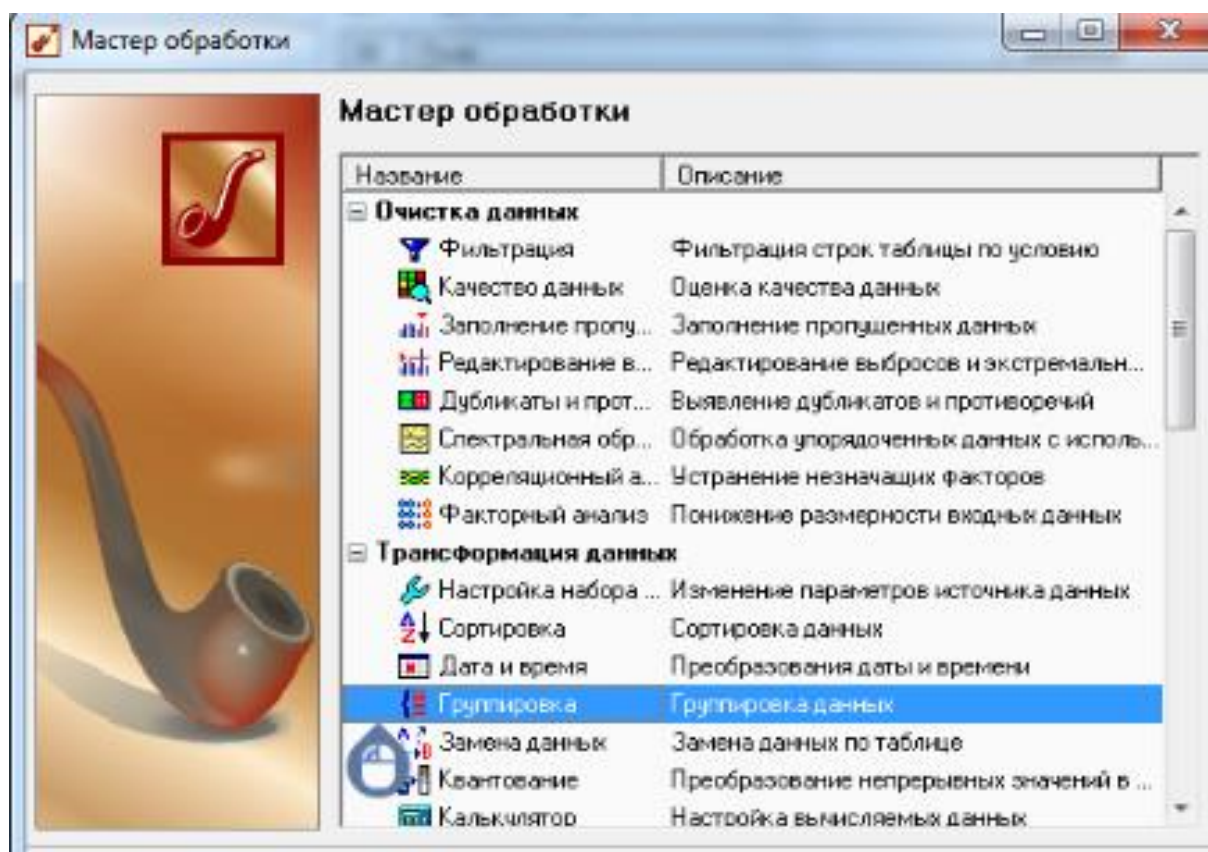


Рисунок 12 – Мастер обработки

На рисунке 13 представлены настройки группировки данных. В нем указывается, какие поля являются измерениями, а какие – фактами.

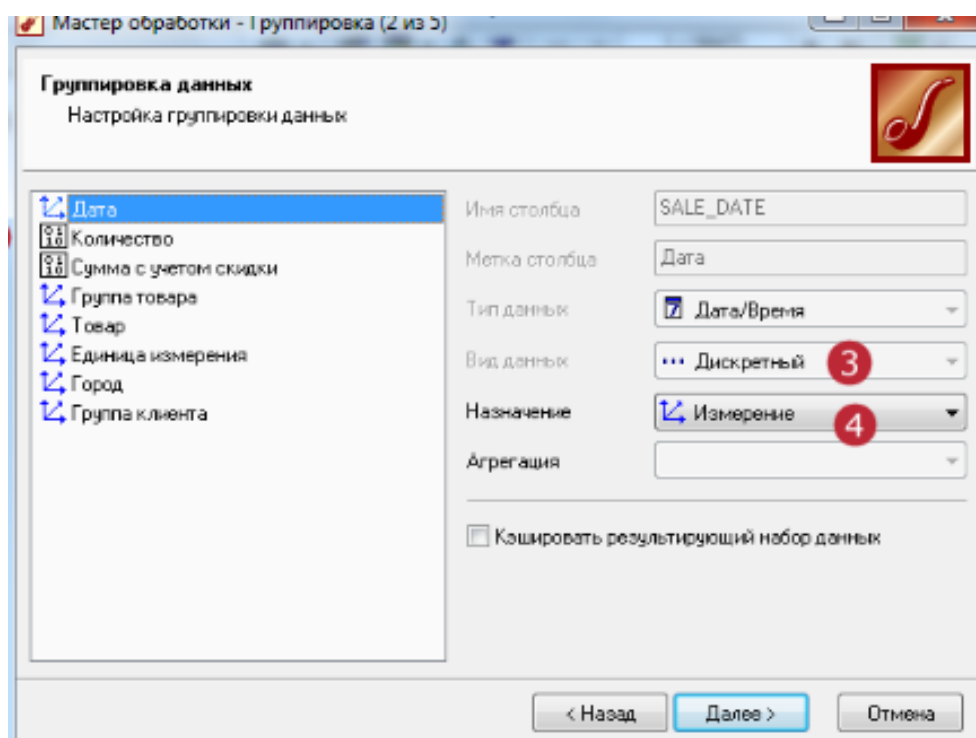


Рисунок 13 – Настройка группировки данных



Поля Группа товара, Единица измерение, Город и Группа клиента не нужны в результирующем наборе данных, т.к. необходимо получить суммарные продажи по дням и товарам. На шаге 2 Мастера обработки отмечаем эти поля как Неиспользуемые.

Для фактов доступно большое число вариантов агрегации, в том числе Медиана, Первый, Последний (элемент в группе). Первый и последний элемент в группе выбирается в соответствии с естественным порядком, в котором эти элементы следуют в исходном наборе данных.

Для категориальных полей:

- в качестве функций агрегации нельзя указать различные статистические функции: Среднее, Медиана и т.п.;
- Максимум (минимум) двух строк рассчитывается посимвольно сравнением;
- Количество – это число агрегируемых значений фактов для каждой комбинации измерений.

Для данной задачи для фактов необходима функция агрегации Сумма, предлагаемая по умолчанию.

Сформировался набор данных по продажам товаров по дням (рисунок 14).

Дата	Товар	Количество	Сумма с учетом скидки
01.03.2004	Банбук половина ствола, диаметр 30-40 мм, 2 м.	42	3290,28
01.03.2004	Болт 10x20 цинк, шестигранная головка, 8 штук, 8	126	2764,44
01.03.2004	Болт 6x80 цинк, шестигранная головка, 3 штуки, 3	112	1245,0144
01.03.2004	Болт 6x90 цинк, шестигранная головка, 6 штук, 6	120	2115,6
01.03.2004	Болт 8x100 цинк, шестигранная головка, 2 штуки, 2	248	3366,22
01.03.2004	Болт 8x25 цинк, шестигранная головка, 8 штук, 8	240	4125,42
01.03.2004	Болт 8x40 цинк, шестигранная головка, 8 штук, 8	120	2501,16
01.03.2004	Болт 8x50 цинк, шестигранная головка, 6 штук, 7	232	4470,6112
01.03.2004	Болт 8x60 цинк, шестигранная головка, 6 штук, 8	112	2383,5616
01.03.2004	Болт 8x70 цинк, шестигранная головка, 6 штук, 9	187	4742,32
01.03.2004	Болт 8x90 цинк, шестигранная головка, 4 штук, 4	120	1669,2
01.03.2004	Болт анкерный 10x125 с гайкой, 5 штук, 7 категор	67	1278,092
01.03.2004	Бордюр KERABEST Alicante 071 25x6, шт	40	1214,4
01.03.2004	Бордюр ИТАЛБАШКЕРАМИКА Брензи шеппанско	40	780
01.03.2004	Бордюр ИТАЛБАШКЕРАМИКА Букет Башкирии 20	40	794,4
01.03.2004	Бордюр ИТАЛБАШКЕРАМИКА Текстиль голубой	40	907,2
01.03.2004	Бордюр ИТАЛБАШКЕРАМИКА Яшма 20x3,5, 1 сор	40	931,2
01.03.2004	Бордюр КЕРАМИН Соната 4 200x35, шт	40	483,74
01.03.2004	Бордюр ШАХТИНСКАЯ ПЛИТКА Муаре 200x28, б	40	781,66
01.03.2004	Бордюр ШАХТИНСКАЯ ПЛИТКА Муаре 200x58, б	40	893
01.03.2004	Бордюр ШАХТИНСКАЯ ПЛИТКА Охота 200x58, шт	40	893
01.03.2004	Бордюр ШАХТИНСКАЯ ПЛИТКА Северина 200x58	40	940
01.03.2004	Бордюр ШАХТИНСКАЯ ПЛИТКА Северина 200x58	40	940
01.03.2004	Бордюр ШАХТИНСКАЯ ПЛИТКА Флейта 200x28, i	40	822,8
01.03.2004	Бордюр ШАХТИНСКАЯ ПЛИТКА Флейта 200x58, i	40	893
01.03.2004	Венецианская штукатурка SENIDECO STUC Acrylic	45	94479,6375
01.03.2004	Воск бесцветный SENIDECO Cire Stuc acrilique, 1 л	56	62828,08
01.03.2004	Воск бесцветный SENIDECO Finish l'acrilique, 0,5 л	49	19001,5434

Рисунок 14 – Итоговый набор данных по продажам по дням

В данном наборе данных потеряна информация о значениях фактов в разрезе исключенных измерений.

Сделаем еще одну группировку – продажи по сумме и количеству по годам. Для этого понадобится указать единственное измерение – Город, а факты останутся теми же (рисунок 15).

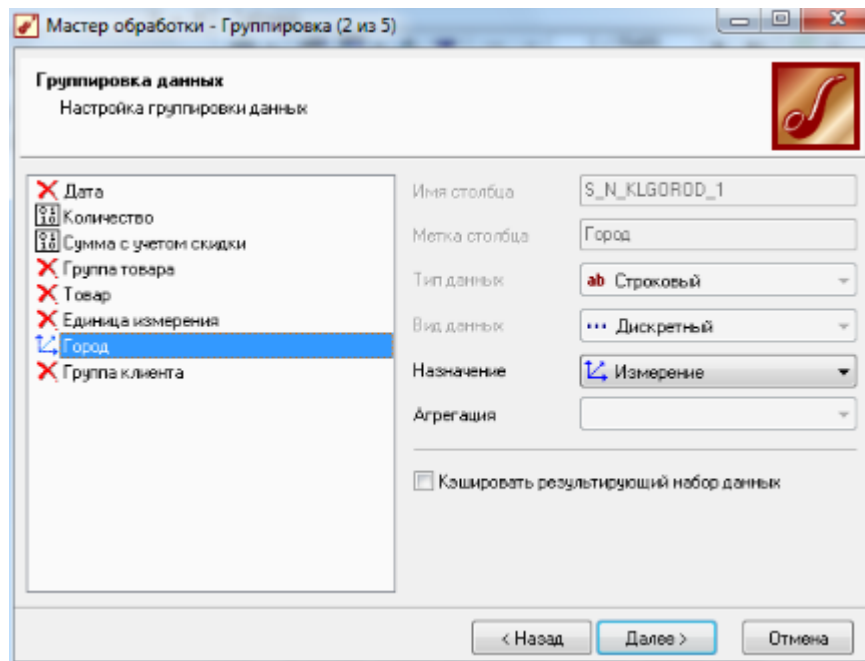


Рисунок 15 – Настройка группировки данных

В итоге получим общие продажи по каждому из 73 городов. Поля Количество и Сумма с учетом скидки отформатированы в визуализаторе Таблица с разделением групп разрядов и округлением.

Существуют ситуации, когда группировка может не иметь измерений. Например, если оставить только факты Количество и Сумма с учетом скидки, получим заданную агрегацию по этим полям (рисунок 16).

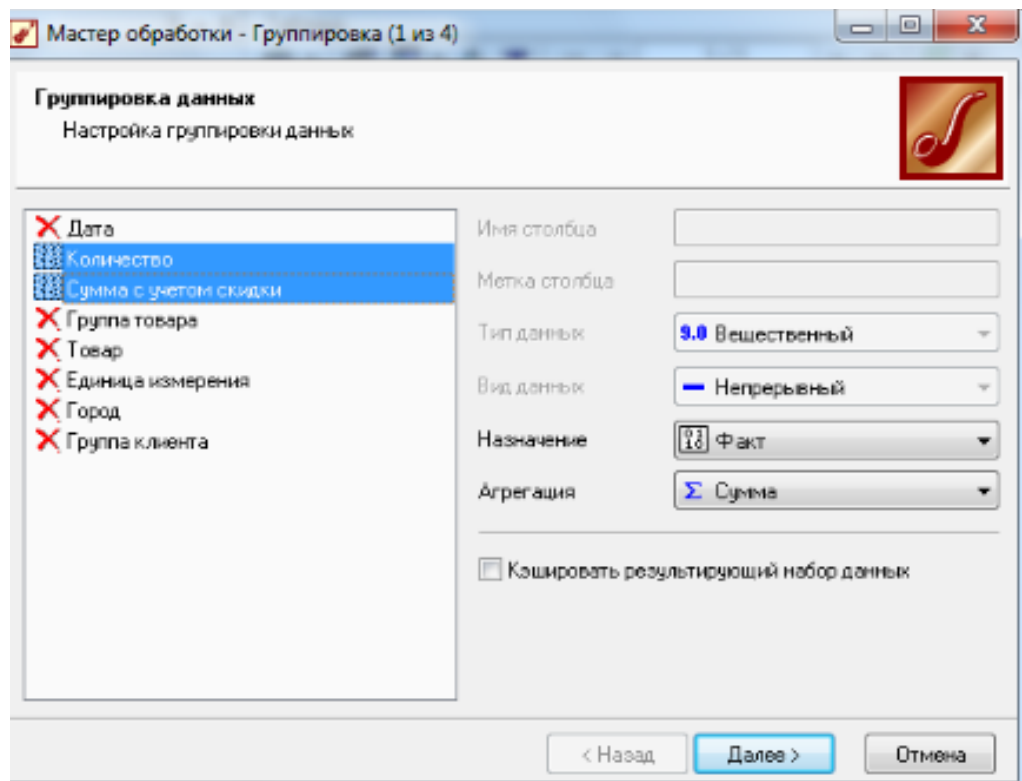
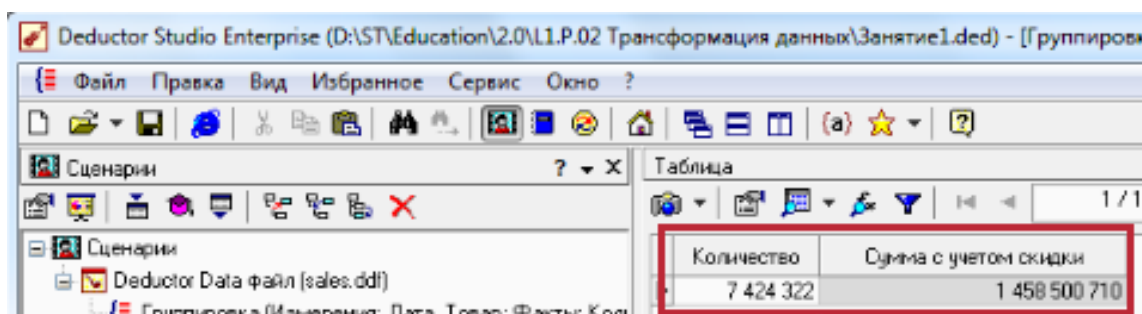


Рисунок 16 – Настройка группировки данных

Получили суммы по полям Количество и Сумма с учетом скидки (рисунок 17).



Количество	Сумма с учетом скидки
7 424 322	1 458 500 710

Рисунок 17 – Результат группировки данных

Импортируем второй файл – plan.ddf. В нем содержится информация о плане продаж (суммы) по годам.

Пусть нам необходимо для дальнейших расчетов иметь в одном наборе данных суммы прошлых продаж и план продаж. Это можно сделать только при помощи обработчика Слияние с узлом.

План продаж содержит только 63 города, т.е. меньше, чем городов, по которым были продажи.

Присоединять план продаж будем ко второму узлу группировки (по городам). Для этого выделим его и запустим мастер обработки (рисунок 18).

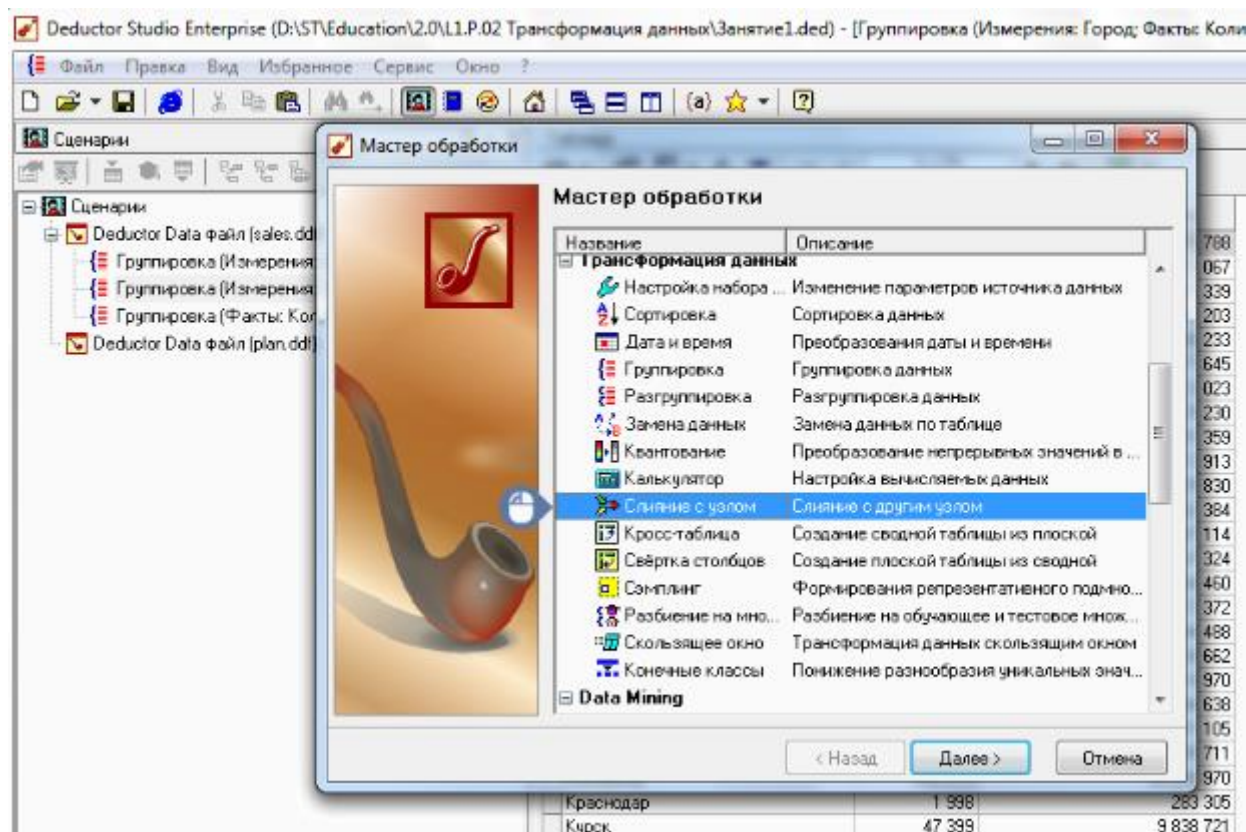


Рисунок 18 – Мастер обработки



В поле «Узел связки» указывается узел сценария, с которым будет происходить слияние. В данном случае присоединяем узел импорта из файла plan.ddf. Выберем тип слияние.

Объединение в данном случае не подходит. Выберем Внутреннее соединение.

На следующем шаге обработчика указывается связь между наборами данных: каким полям из входящего набора соответствуют поля в связанной таблице.

Deductor всегда пытается соединить наборы, если имеются совпадения в именах полей. Команду «Соединить по умолчанию» можно вызвать, открыв меню правой кнопкой мыши.

Если проставленные по умолчанию соответствия не устраивают, в этом же меню нужно вызвать команду Очистить все соответствия.

В нашем случае все правильно: внутреннее соединение должно вестись по полю Город (рисунок 19).

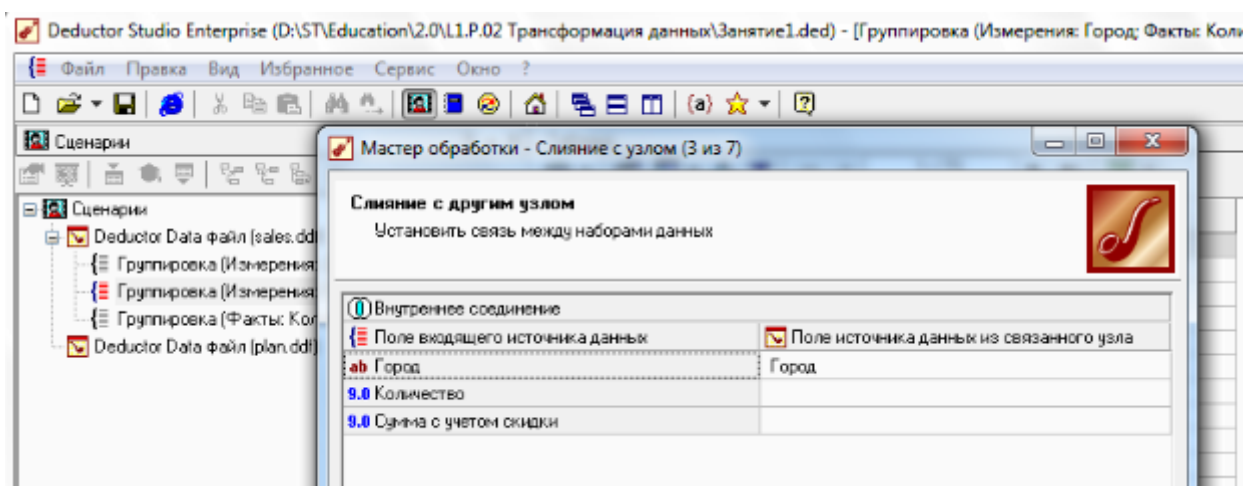


Рисунок 19 – Мастер обработки «Слияние с узлом»

На следующем шаге происходит выбор полей для включения в результирующий набор данных.

В нашем примере это все поля Входящего источника и поле План продаж Источника данных из связанного узла.

В результате после внутреннего соединения получаем набор данных с новым полем плана продаж, то только для тех городов, которые присутствовали в обоих наборах – входящем и связанным с ним источником. Это 62 города (рисунок 20).

Дедуктор Студио Энтерпрайз (D:\ST\Education\2.0\1.1.P.02 Трансформация данных\Занятие1.ded) - [Внутреннее соединение (Deductor Data файл (plan.ddf))]

Файл Правка Вид Избранное Сервис Окно ?

Сценарии

- Дедуктор Data файл (sales.ddf)
  - Группировка (Измерения: Дата, Товар; Факты: Кол
  - Группировка (Измерения: Город; Факты: Количес
  - Внутреннее соединение (Deductor Data файл (plan
  - Группировка (Факты: Количество, Сумма с учетом с
- Дедуктор Data файл (plan.ddf)

Таблица

22 / 62

Город	Количество	Сумма с учетом скидки	План продаж+
Архангельск	27 475	5 415 788	7 326 225
Астрахань	18 734	2 913 067	3 011 748
Балашика	43 631	9 823 339	14 719 099
Барнаул	95 211	16 145 203	17 749 851
Белгород	3 883	580 233	604 537
Владивосток	16 006	2 323 023	2 567 965
Владимир	351 686	71 025 230	106 313 094
Волгоград	63 225	11 381 359	13 932 240
Вологда	88 032	13 272 913	14 593 668
Воронеж	81 752	19 325 830	23 821 753
Екатеринбург	10 436	1 882 384	2 094 496
Зеленоград	144 000	26 587 114	36 583 861
Иваново	141 437	30 147 324	32 823 203
Ижевск	18 365	2 664 460	3 095 749
Иркутск	3 669	572 372	767 634
Казань	57 217	9 894 488	12 539 176
Калуга	34 001	8 930 662	13 191 520
Кемерово	53 492	12 118 970	14 520 168
Кострома	283 018	55 849 970	75 333 265
Краснодар	1 998	283 305	332 416
Курск	47 399	9 838 721	10 642 954
Магайск	25 741	5 660 139	7 106 619
Москва	2 499 028	476 055 694	599 309 701
Мурманск	15 211	2 003 888	2 703 832

Рисунок 20 – Результат операции Внутреннее соединение

Добавим такой же узел слияния, но используя Внешнее левое соединение.

Мы получим 73 записи – ровно столько, сколько было в наборе, содержащим продажи по городам. Для тех городов, где не нашлось плана продаж из файла plan.ddf, стоит NULL-значение (пусто).

Если выбрать третий вариант, Внешнее правое соединение, то получим 63 записи. Из файла плана добавится город Ханты-Мансийск, но т.к. по нему не было данных в таблице продаж, то соответствующие поля записи будут пустыми (рисунок 21).

Deductor Studio Enterprise (D:\ST\Education\2.0\U1.P.02 Трансформация данных\Занятие1.ded) - (Внешне

Файл Правка Вид Избранное Сервис Окно ?

Таблица

63 / 63

Город	Количество	Сумма с учетом скидки	План продаж+	Город+
Пермь	105429	19259598,13	23457887	Пермь
Петрозаводск	33719	6893788,11	9227014	Петрозаводск
Псков	28432	3932238,08	4428963	Псков
Пушкино	45249	10540524,15	13072570	Пушкино
Ростов-на-Дону	42779	9270006,7825	11519376	Ростов-на-Дону
Рыбинск	40006	6481490,4225	9615710	Рыбинск
Рязань	16363	3545404,52	4528863	Рязань
Самара	40605	8051614,33	8413562	Самара
Санкт-Петербург	4629	547945,99	741152	Санкт-Петербург
Саратов	17710	3167736,48	3385412	Саратов
Смоленск	1446	202410,43	246858	Смоленск
Сочи	66483	15126137,85	21606879	Сочи
Стерлитамак	10243	1811245,53	1951287	Стерлитамак
Сургут	20575	3493692,61	3656424	Сургут
Сыктывкар	19814	2744855,86	3501951	Сыктывкар
Тамбов	3684	695619,16	872214	Тамбов
Тверь	37495	5308947,67	5382037	Тверь
Тольятти	151669	33823757,0074	45307079	Тольятти
Томск	38801	5651034,07	6293407	Томск
Тула	39790	6585681,40749999	9274882	Тула
Улан-Удэ	15618	2251455,27	2494254	Улан-Удэ
Ульяновск	11192	1714489,7	2394201	Ульяновск
Уссурийск	9688	1435470,84	1753330	Уссурийск
Уфа	82112	20805622,08	21074959	Уфа
Хабаровск	29818	6008245,11	6630504	Хабаровск
Челябинск	43833	8275200,75	12331484	Челябинск
Чита	7677	1449229,11	1809462	Чита
Ярославль	28827	7313518,84	9385302	Ярославль
			576400	Ханты-Мансийск

Рисунок 21 – Результат Внутреннего правого соединения

Необходимость в использовании Полного соединения возникает крайне редко. Типичная ситуация - требуется присоединить ко всем записям набора данных какое-то одно значение, содержащееся в другом узле. Рассмотрим пример присоединения общего значения Сумма с учетом скидки, полученного ранее группировкой по всем записям и равное числу 1 458 500 710.

В Полном соединении допускается не указывать ни одной связи.

Получили требуемый результат. При слиянии на последнем шаге переименовали присоединяемое поле в Общая сумма (рисунок 22).

Город	Количество	Сумма с учетом скидки	Общая сумма
Архангельск	27475	5 415 798	1 458 500 710
Астрахань	18734	2 913 067	1 458 500 710
Балашиха	43631	9 823 339	1 458 500 710
Барнаул	95211	16 145 203	1 458 500 710
Белгород	3883	580 233	1 458 500 710
Владивосток	16006	2 323 023	1 458 500 710
Владимир	351686	71 025 230	1 458 500 710
Волгоград	63225	11 381 359	1 458 500 710
Вологда	88032	13 272 913	1 458 500 710
Воронеж	81752	19 325 830	1 458 500 710
Екатеринбург	10436	1 882 384	1 458 500 710
Зеленоград	144000	26 587 114	1 458 500 710
Иваново	141437	30 147 324	1 458 500 710

Рисунок 22 – Результат Полного объединения

### Пример №2 «Узлы Дата и время и Скользящее окно».

В данном примере используются наборы данных sales.ddf и calendar.ddf (находятся в папке «К лабораторной работе №5»).

**Самостоятельно сделайте набор данных с продажами по дням по всем городам.**

В итоге получаем временные ряды продаж товаров по дням.

Необходимо получить продажи по месяцам года. Нужно сделать группировку, но поля, по которому нужно группировать – месяц года – нет. Его можно получить из поля Дата продажи с помощью обработчика Дата и Время (рисунок 23).

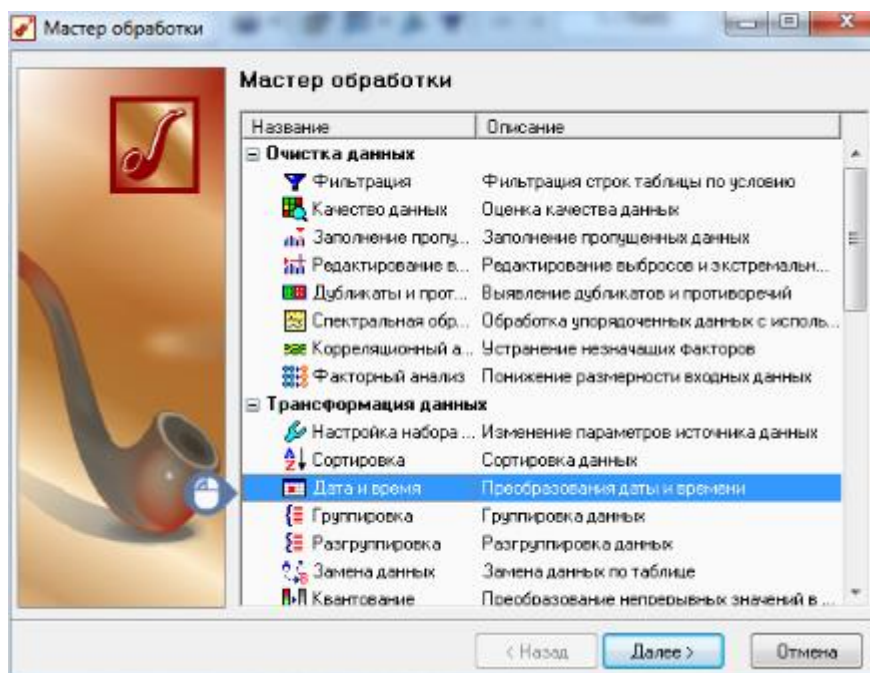


Рисунок 23 – Мастер обработки

Суть такого преобразования заключается в том, что на основе столбца с информацией о дате/времени формируются один или несколько столбцов, в которых указывается, к какому заданному интервалу времени принадлежит строка данных.

По умолчанию первому полю из списка с типом Дата/время устанавливается назначение – Используемое и назначается тип разбиения Год+Месяц (рисунок 24).

Преобразование даты и времени  
Преобразование даты и времени

☒ Дата  
☐ Товар  
☐ Количество  
☐ Сумма с учетом скидки

Имя столбца: SALE DATE  
Назначение: ☒ Используемое

Разбиение	Тип данных		
	Дата	Число	Строка
Год + Квартал	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Год + Месяц	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Год + Неделя	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Год + День	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Год	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Квартал	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Месяц	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Неделя	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
День года	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
По умолчанию	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

☐ Обработать даты по ISO 8601

< Назад    Далее >    Отмена

Рисунок 24 – Мастер обработки «Дата и время»

Дополнительно выделим из поля Дата продажи наименование месяца и номер недели продажи. Получаемые поля в результате преобразования должны быть следующего типа: месяц – Строковый; неделя – Числовой.

В результате появились три новых поля (рисунок 25).



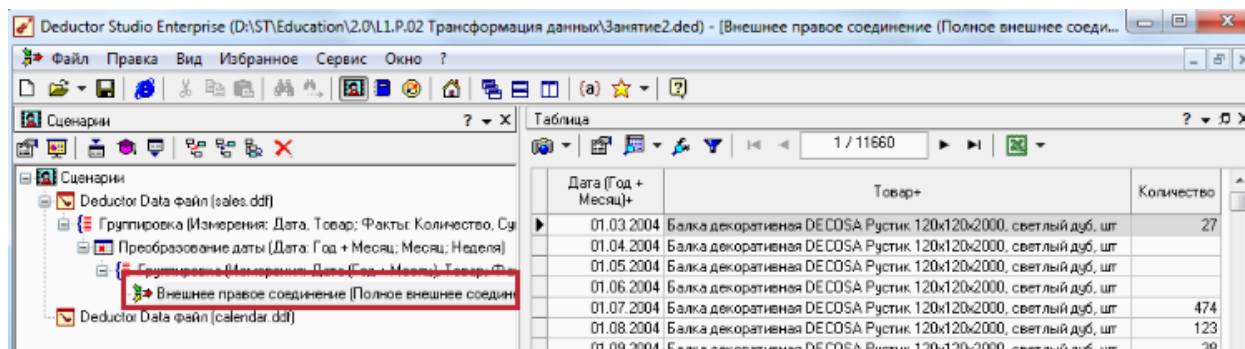


По каждому товару нужно получить таблицу, где каждое значение наблюдения должно быть дополнено значениями наблюдений за 2 прошлых месяца и за 1 будущий месяц. Эта задача решается узлом Скользящее окно.

Обозначим проблему: в некоторые месяцы продаж определенных товаров не было. История продаж начинается с марта 2004 года и данные по июню, октябрю и декабрю отсутствуют. Поэтому сначала нужно заполнить их нулями.

Обычно это делается при помощи специального справочника – «Календаря», где для каждого товара сформирован полный ряд дат, в данном случае Год+Месяц. В нашем случае это период с марта по декабрь 2004 года. Импортируйте в сценарий файл calendar.ddf.

Внешним правым соединением с календарем в поле Количество получим пустые записи по тем месяцам, в которых не было продаж (рисунок 27).



Дата (Год + Месяц)+	Товар+	Количество
01.03.2004	Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	27
01.04.2004	Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	
01.05.2004	Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	
01.06.2004	Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	
01.07.2004	Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	474
01.08.2004	Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	123
01.09.2004	Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	38

Рисунок 27 – Результат Внешнего правого соединения

Дальше потребуется узел Замена данных для замены пустых значений поля Количество на нули (рисунок 28).

Таблица			
23 / 11680			
Дата (Год + Месяц)+	Товар+	Количество	Количество_REPLACE
01.03.2004	Балка декоративная DECOSA Рустик	27	27
01.04.2004	Балка декоративная DECOSA Рустик		0
01.05.2004	Балка декоративная DECOSA Рустик		0
01.06.2004	Балка декоративная DECOSA Рустик		0
01.07.2004	Балка декоративная DECOSA Рустик	474	474
01.08.2004	Балка декоративная DECOSA Рустик	123	123
01.09.2004	Балка декоративная DECOSA Рустик	38	38
01.10.2004	Балка декоративная DECOSA Рустик		0
01.11.2004	Балка декоративная DECOSA Рустик	184	184
01.12.2004	Балка декоративная DECOSA Рустик		0
01.03.2004	Балка декоративная DECOSA Рустик		0
01.04.2004	Балка декоративная DECOSA Рустик	81	81
01.05.2004	Балка декоративная DECOSA Рустик		0
01.06.2004	Балка декоративная DECOSA Рустик		0
01.07.2004	Балка декоративная DECOSA Рустик	41	41
01.08.2004	Балка декоративная DECOSA Рустик		0
01.09.2004	Балка декоративная DECOSA Рустик		0
01.10.2004	Балка декоративная DECOSA Рустик	265	265
01.11.2004	Балка декоративная DECOSA Рустик	184	184
01.12.2004	Балка декоративная DECOSA Рустик		0
01.03.2004	Балка декоративная DECOSA Рустик		0
01.04.2004	Балка декоративная DECOSA Рустик	105	105
01.05.2004	Балка декоративная DECOSA Рустик		0
01.06.2004	Балка декоративная DECOSA Рустик		0
01.07.2004	Балка декоративная DECOSA Рустик		0
01.08.2004	Балка декоративная DECOSA Рустик		0

Рисунок 28 – Результат Замены данных

Теперь настройкой набора данных уберем ненужные поля, а затем поставим фильтр на какой-либо один товар и отсортируем по возрастанию даты.

Скользящее окно нужно делать для каждого товара по отдельности. Значит, это групповая обработка. Поэтому реализована ветвь сценария на 1 товаре, а затем распространится на все остальные.

Обработчик Скользящее окно преобразовывает исходную структуру таблицы в новую, добавляя столбцы смещенных значений записей.

В окне настройки параметров обработчика выбираем поле Количество и изменяем его назначение на Используемое. Для поля откроются параметры настройки скользящего окна.

Зададим параметры скользящего окна на основе двух предыдущих месяцев и одного будущего:

- Глубина погружения = 2;
- Горизонт прогнозирования = 1;
- Флаг Оставлять неполные записи – выключен.

В результате в наборе данных появились три новых поля (рисунок 29).



Товар	Дата (Год + Месяц)+	Количество-2	Количество-1	Количество	Количество+1
Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	01.06.2004	27	0	0	0
Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	01.06.2004	0	0	0	474
Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	01.07.2004	0	0	474	123
Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	01.08.2004	0	474	123	38
Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	01.09.2004	474	123	38	0
Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	01.10.2004	123	38	0	184
Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	01.11.2004	38	0	184	0

Рисунок 29 – Результирующий набор данных

Количество новых столбцов в результирующем наборе данных равно сумме параметров Глубина погружения и Горизонт планирования.

Метки этих полей проставляются по правилу: <Метка>-/+<Число>, Знак «-»/«+» показывает на направления смещение назад/вперед. <число> указывает на количество периодов (строк) смещения.

Если включить флаг Оставлять неполные записи, то получится такой результат (рисунок 30).

Товар	Количество-2	Количество-1	Количество	Количество+1	Дата (Год + Месяц)+
Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт				27	
Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт			27	0	01.03.2004
Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт		27	0	0	01.04.2004
Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	27	0	0	0	01.05.2004

Рисунок 30 – Результат установки флага «Оставлять неполные записи»

Скользящее окно может потребоваться не только в задачах подготовки временных рядов для прогнозирования, но и при создании каких-либо нетривиальных пользовательских отчетов, например, вычисление времени между событиями в строках.

Последнее, что осталось сделать – применить групповую обработку, распространив скользящее окно на все товары. «Все товары» находятся перед узлом фильтрации.

Товар будет группой при обработке данных (рисунок 31).

Метка столбца	Имя столбца
<input checked="" type="checkbox"/> Товар	ARTICLE_NAME
<input type="checkbox"/> Количество	9.0 COUNT
<input type="checkbox"/> Дата (Год + Месяц)+	DATE_i

Рисунок 31 – Определение групп обработки

В качестве начального узла групповой обработки данных выбираем «Сортировка: Дата (Год+Месяц)».

В качестве конечного узла групповой обработки данных выбираем «Скользящее окно (Количество)».

В результате получили 8162 записи (рисунок 32) – набор данных готов к дальнейшей обработке, например, построению прогноза.

Товар	Количество-2	Количество-1	Количество	Количество+1	Дата (Год + Месяц)+
▶ Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	27	0	0	0	01.05.2004
Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	0	0	0	474	01.06.2004
Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	0	0	474	123	01.07.2004
Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	0	474	123	38	01.08.2004
Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	474	123	38	0	01.09.2004
Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	123	38	0	184	01.10.2004
Балка декоративная DECOSA Рустик 120x120x2000, светлый дуб, шт	38	0	184	0	01.11.2004
Балка декоративная DECOSA Рустик 60x90x2000, темный дуб, шт	0	81	0	0	01.05.2004
Балка декоративная DECOSA Рустик 60x90x2000, темный дуб, шт	81	0	0	41	01.06.2004
Балка декоративная DECOSA Рустик 60x90x2000, темный дуб, шт	0	0	41	0	01.07.2004
Балка декоративная DECOSA Рустик 60x90x2000, темный дуб, шт	0	41	0	0	01.08.2004
Балка декоративная DECOSA Рустик 60x90x2000, темный дуб, шт	41	0	0	265	01.09.2004
Балка декоративная DECOSA Рустик 60x90x2000, темный дуб, шт	0	0	265	184	01.10.2004
Балка декоративная DECOSA Рустик 60x90x2000, темный дуб, шт	0	265	184	0	01.11.2004
Балка декоративная DECOSA Рустик 60x90x3000, темный дуб, шт	0	105	0	0	01.05.2004
Балка декоративная DECOSA Рустик 60x90x3000, темный дуб, шт	105	0	0	0	01.06.2004
Балка декоративная DECOSA Рустик 60x90x3000, темный дуб, шт	0	0	0	0	01.07.2004
Балка декоративная DECOSA Рустик 60x90x3000, темный дуб, шт	0	0	0	345	01.08.2004
Балка декоративная DECOSA Рустик 60x90x3000, темный дуб, шт	0	0	345	265	01.09.2004
Балка декоративная DECOSA Рустик 60x90x3000, темный дуб, шт	0	345	265	0	01.10.2004
Балка декоративная DECOSA Рустик 60x90x3000, темный дуб, шт	345	265	0	0	01.11.2004
Балка декоративная DECOSA Рустик 60x90x4000, темный дуб, шт	27	147	244	84	01.05.2004
Балка декоративная DECOSA Рустик 60x90x4000, темный дуб, шт	147	244	84	375	01.06.2004
Балка декоративная DECOSA Рустик 60x90x4000, темный дуб, шт	244	84	375	123	01.07.2004
Балка декоративная DECOSA Рустик 60x90x4000, темный дуб, шт	84	375	123	114	01.08.2004
Балка декоративная DECOSA Рустик 60x90x4000, темный дуб, шт	375	123	114	0	01.09.2004
Балка декоративная DECOSA Рустик 60x90x4000, темный дуб, шт	123	114	0	292	01.10.2004

Рисунок 32 – Результирующий набор данных

### Пример №3 «Квантование».

В данном примере используется набор данных Регионы.txt (находится в папке «К лабораторной работе №5»).

Узел Квантование осуществляет разбиение диапазона числовых значений на указанное количество интервалов (рисунок 33).

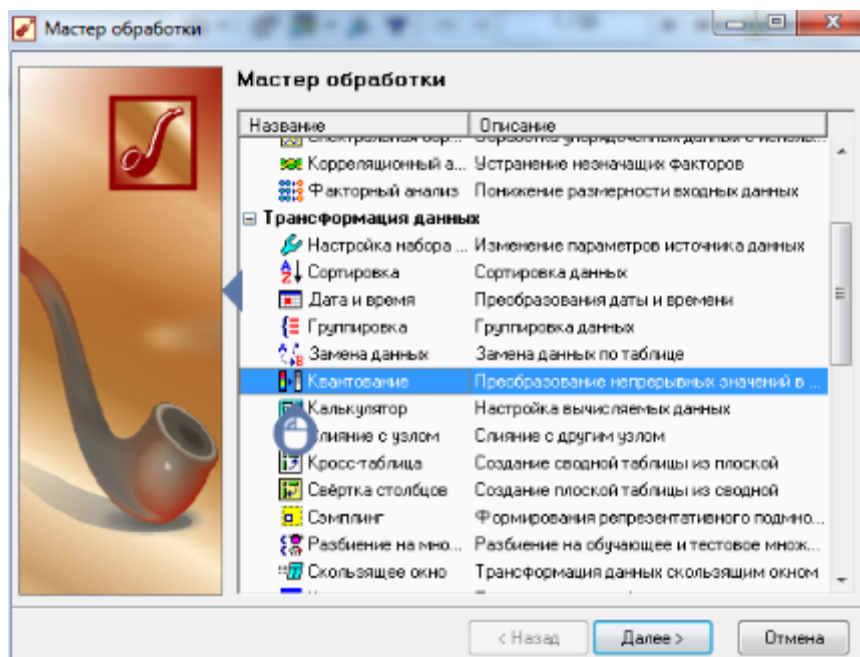


Рисунок 33 – Мастер обработки

На втором шаге открывается список полей набора данных. Поле может быть использовано для квантования значений, если выполнены условия:

- тип поля числовой (целый или вещественный);
- стандартное отклонение столбца не равно нулю, т.е. поле содержит хотя бы 2 уникальных значения.

В противном случае поле помечается как Непригодное.

Выполним квантование для поля Численность населения (тыс.чел). Назначение поля должно быть Используемое.

Доступны два способа разбиения – По интервалам и По квантилям:

- при интервальном способе диапазон исходных значений разбивается на равные интервалы;
- при квантильном – интервалы выбираются таким образом, чтобы в каждый из них попадало одинаковое количество значений.

Выбираем вариант По интервалам.

Зададим количество интервалов 5. В списке Значение выберем вариант представления результатов квантования, т.е. по какому правилу будут формироваться значения интервалов.

Оставим пункт Автоматическая метка (рисунок 34).

- Номер интервала – отображаются номера интервалов, в которые попали значения;
- Нижняя граница – отображаются нижние границы интервалов;
- Верхняя граница – отображаются верхние границы интервалов;
- Середина интервала – отображаются средние значения интервалов;
- Метка интервала – отображаются пользовательские метки, которые нужно будет знать;

– Автоматическая метка – отображаются метки, которые формируются автоматически исходя из границ интервала.

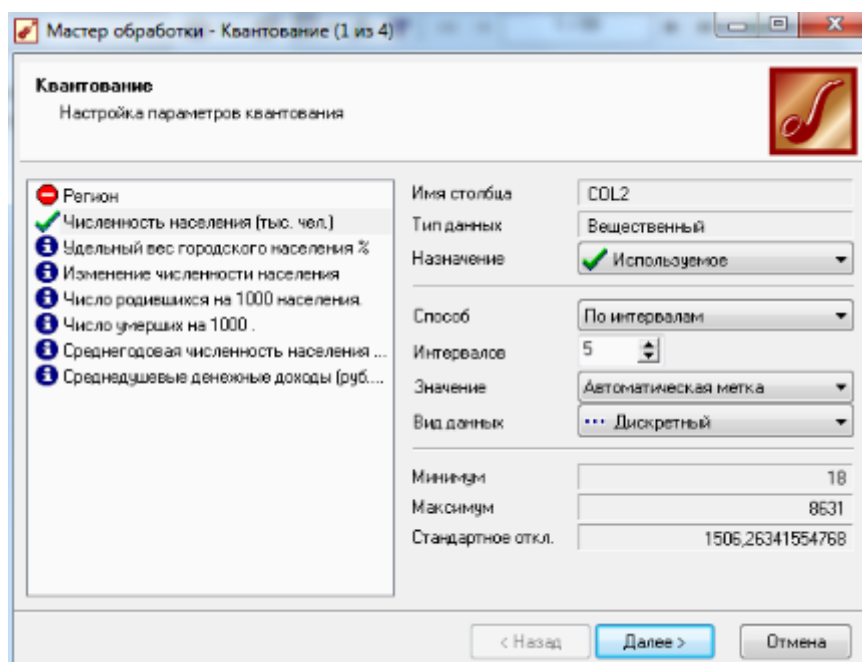


Рисунок 34 – Мастер обработки – Квантование

На следующем шаге настраиваются границы интервалов. Поскольку выбрана автоматическая метка и способ разбиения по интервалам, то мастер рассчитает интервалы одинаковой длины и сформирует для них соответствующие метки.

При прогоне через этот узел новых данных или копировании узла и его присоединении к другим данным, интервалы будут пересчитаны согласно новым данным поля.

Если вручную исправить любую границу, то интервалы будут зафиксированы, и останутся такими для любого обрабатываемого столбца. При этом в заголовке таблицы к слову Интервалы добавится слово «(изменены)».

В результате квантования «старое» поле заменится на квантованное.

#### **Пример №4 «Кросс-таблица и Свертка столбцов».**

В данном примере используется набор данных Потребление электрической энергии.txt (находится в папке «К лабораторной работе №5»).

Перед нами временной ряд потребления электроэнергии. Необходимо выполнить транспонирование этого набора данных так чтобы каждый объект располагался бы в отдельном столбце.

Самый простой и быстрый способ это сделать – применить узел Кросс-таблица (рисунок 35).

Обработчик Кросс-таблица создает сводную таблицу из плоской. Значения полей переносится в заголовки столбцов.

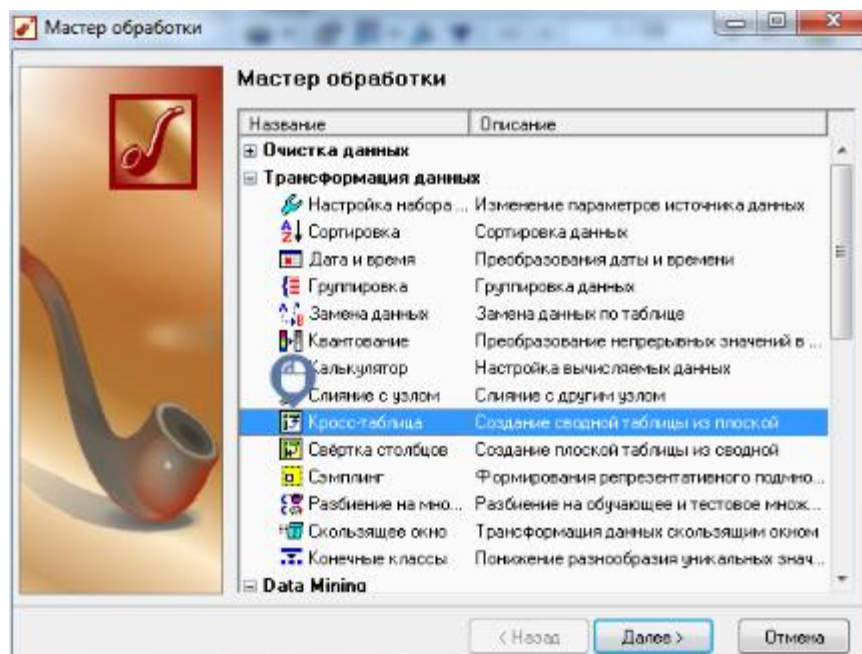


Рисунок 35 – Мастер обработки

На втором шаге мастера задается структура нового набора данных: какие из полей исходной таблицы станут колонками, а какие – столбцами.

Искусственное поле Количество присутствует всегда, и может быть добавлено только в группу Факты. В нем будет подсчитано, сколько раз в исходном наборе данных встречается каждая комбинация из колонок и строк.

Зададим структуру нового набора данных. Поле Объект поместим в группу Колонки для того чтобы потребление электроэнергии для каждого объекта выводилось в отдельном столбце. Поле Дата – в группу Строки, а Потребление электроэнергии, кВт час – в Факты.

На третьем шаге открывается окно настройки агрегации фактов (рисунок 36).

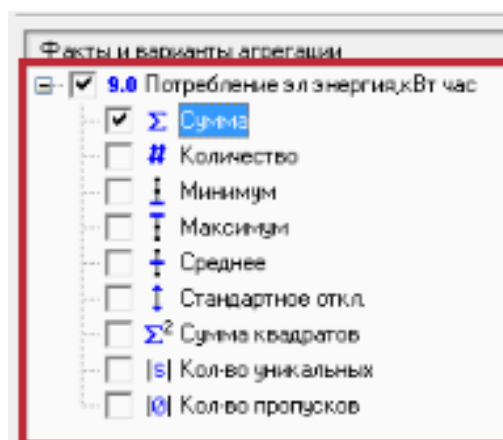


Рисунок 36 – Настройка фактов

Если записей по конкретному объекту на конкретную дату несколько, то необходимо выбрать как их объединить в одну: просуммировать, взять среднее, минимум и т.д. Для категориальных полей вариантов агрегации будет меньше. Агрегация аналогична той, что уже встречалась в узле Группировка.

Оставим вид агрегации, предложенный по умолчанию – Сумма. В данном случае без разницы что выбрать – сумму, минимум, максимум, среднее – в исходном наборе данных по каждому объекту и дате только 1 запись.

Следующий шаг мастера – настройка параметров измерений в колонках. Деактивируем флаг Пропущенные и установим флаг Прочие.

Первоначальный столбец Потребление эл энергии кВт час разделится на четыре столбца (рисунок 37): в первых трех – уникальные наименования объектов плюс столбец Прочее, который мы включили в набор данных специальным флагом.

Объект 1	Объект 2	Объект 3	«Прочее»
25668			
23292			
25155			
24228			
21510			
18513			
20079			
21951			
31212			
32688			
30438			
28764			
27387			
27171			
23037,669		128933,1	
		111480	
20361,525		113947,5	
21972,3		122970	
20659,725		115627,5	
22602,975		126502,5	
26337,15		147397,5	
26279,25		147067,5	
36532,425		204465	
29880,45		167227,5	
32113,125		179730	
30285,675		169507,5	
23484,975		131437,5	
26381,625	185127,5	147652,5	
18588,675	225127,5	104205	

Рисунок 37 – Результат применения узла Кросс-таблица

Здесь находится NULL-значение («пусто»), т.к. в исходном наборе данных нет ни одной записи о потреблении электроэнергии в мае 2009 года по Объекту 1.

Создадим второй узел кросс-таблицы с такими же настройками, но включенным флагом скользящих уникальных значений. Для этого скопируем уже существующий узел и перенастроим его.

Поднимем флаг «Скользящие уникальные значения».



Искусственно создадим ситуацию, при которой в наборе данных появляются новые уникальные значения. Для этого при помощи узла Замена данным заменим значение поля Объект «Объект 3» на «Объект 4».

Теперь последовательно двумя скриптами «прогоним» измененный набор данных через кросс-таблицы: сначала без флага со скользящими значениями, а потом – с ним.

В результате получилось два набора данных (рисунок 38).

**Скрипт (18-18): ("Кросс-таблица (с флагом Прочее)")**

Дата	Объект 1	Объект 2	Объект 3	<Прочее>
01.02.2009	27387			
01.03.2009	27171			
01.04.2009	23037,669			128933,1
01.05.2009				111480
01.06.2009	20361,525			113947,5
01.07.2009	21972,3			122970
01.08.2009	20659,725			115627,5
01.09.2009	22602,975			126502,5
01.10.2009	26337,15			147387,5

**Скрипт (20-20): ("Кросс-таблица (с флагом Скользящие уникальные значения)")**

Дата	Объект 1	Объект 2	Объект 4
01.01.2009	28764		
01.02.2009	27387		
01.03.2009	27171		
01.04.2009	23037,669		128933,1
01.05.2009			111480
01.06.2009	20361,525		113947,5
01.07.2009	21972,3		122970
01.08.2009	20659,725		115627,5

Рисунок 38 – Результат выполнения скриптов

В первом случае все значения Объекта 4 были перенесены в поле Прочее, т.к. на момент настройки узла данного уникального значения в наборе не было, а ранее созданное поле Объект 3 осталось пустым.

Во втором случае колонки скорректировались на основе уникальных значений поля Объект и ранее созданное поле Объект 3 заменилось на Объект 4.

Вернем набор данных, преобразованный с помощью узла Кросс-таблица, в первоначальный вид. Это обратная задача, и она решается узлом Свертка столбцов (рисунок 39).

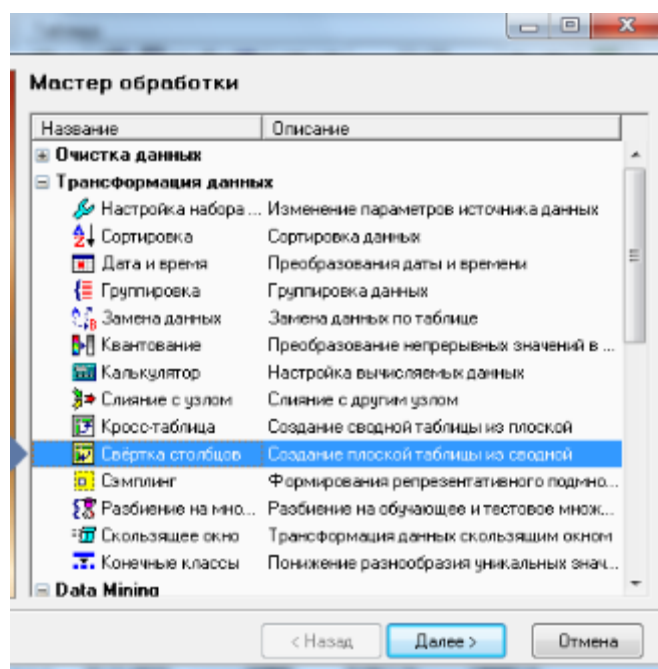


Рисунок 39 – Мастер обработки

Обработчик Свертка столбцов из сводной таблицы делает плоскую. При этом заголовки полей переносится в значения строк и столбцов.

На втором шаге открывается окно формирования структуры нового набора данных. В нем определяется назначение полей.

Зададим структуру нового набора данных. Поле Дата оставим без изменения (Информационные), а значения потребления электроэнергии по объектам объединенным в один столбец (Транспонируемые).

Следующий шаг мастера открывает окно задания меток формируемым столбцам. По умолчанию предлагаются метки с именем Заголовки и Значения (рисунок 40).



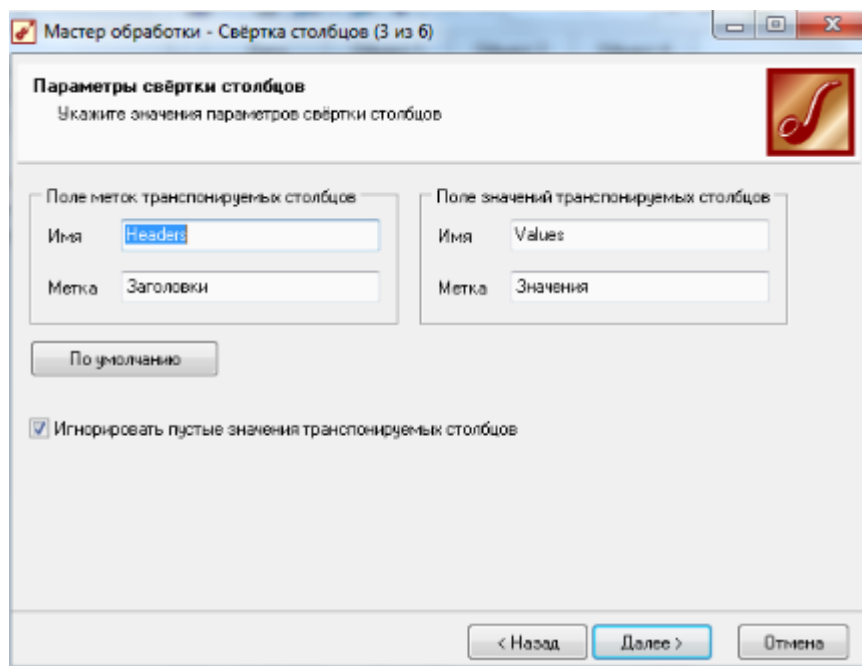


Рисунок 40 – Параметры свертки столбцов

Изменим метки: Заголовки на Объекты; Значения на Потребление эл энергии, кВт час.

В результате получен исходный набор данных.

#### Задания для самостоятельной работы:

1. Получите файл calendar.ddf самостоятельно из файла sales.ddf при помощи небольшого сценария.
2. Выполните импорт текстового файла Потребление электрической энергии.txt. Рассчитайте прирост потребления электроэнергии по сравнению с предыдущим месяцем.
3. Изучите в калькуляторе функцию Data(). Рассчитайте прирост потребления электроэнергии с помощью данной функции.
4. Выполните квантование для текстового файла Регионы.txt используя 2 способа: 1) автоматическая метка и 2) измените какие-либо границы автоматических меток. Получите 2 узла квантования.
5. Используя узел Скрипт, примените квантование к полю Среднегодовая численность населения занятого в экономике используя для поля Численность населения, поочередно к каждому из двух узлов квантования.
6. Дан файл statuses.txt с данными следующей структуры (таблица 1).

Таблица 1 – Структура файла statuses.txt

Заявка.Код	Заявка.Статус	Дата и время статуса
1048	0	20.10.2014 10:02:00
1048	20	20.10.2014 10:22:00

1048	40	20.10.2014 10:23:00
1048	60	20.10.2014 10:42:00
1050	0	21.10.2014 14:51:00
1050	10	21.10.2014 15:04:00
1051	0	21.10.2014 16:00:00
...	...	...

Числовые коды статусов имеют сквозную нумерацию и несут в себе следующую семантику:

- 0 – заявка принята к рассмотрению;
- 10 – заявка не прошла проверку на Этапе 1;
- 20 – заявка прошла проверку на Этапе 1;
- 30 – заявка не прошла проверку на этапе 2;
- 40 – заявка прошла проверку на Этапе 2;
- 50 – заявка не прошла проверку на Этапе 3;
- 60 – заявка прошла проверку на Этапе 3.

Если заявка не проходит проверку на каком-либо этапе, ее жизненный цикл на этом заканчивается.

Требуется сделать преобразования набора данных так, чтобы получить для каждой заявки время ее обработки с момента рассмотрения до решения на Этапе 1. Результирующий набор данных должен иметь следующий вид (таблица 2). Для каждой заявки рассчитывается как время обработки в минутах, так и дискретный аналог времени со следующими интервалами квантования: до 15 мин, 15-29 мин, 30-44 мин, 45-59 мин, от 1 до 2 ч., свыше 2 ч.

Таблица 2 – Результирующий набор данных

Заявка.Код	Время обработки (Минуты)	Время обработки
1048	19	15-29 мин
1050	13	до 15 мин
1051	29	15-29 мин
...	...	...

Условие: при решении задачи запрещается использовать обработчики: **Слияние, Кросс-таблица**. Задача легче всего решается, если использовать **Скользящее окно**.

7. Дан файл sales.txt следующей структуры:

Таблица 3 – Структура файла sales.txt

Дата (Мес)	Товар.Группа	Товар.Код	Клиент.Код	Количество	Сумма
01.03.2014	Группа 2	0008709	0000009	1	9,18
01.03.2014	Группа 2	0008714	0000009	2	18,36

01.03.2014	Группа 1	0010595	0000009	3	48,95
...	...	...	...	...	...

В файле представлена информация о том, какой товар, на какую сумму и в каком количестве был отпущен клиентам.

Требуется:

1 Для каждого товара, который продавался, рассчитать среднюю цену за последние 6 месяцев от имеющихся данных.

2 Для каждого товара, который продавался, рассчитать среднюю цену в каждом месяце и округлить ее до двух знаков после запятой.

3 Для каждого клиента, который делал покупки, рассчитать общую сумму покупок, сделанные им за последние 2 месяца.

Все результаты должны быть округлены до 2-х знаков после запятой.

Внимание! Типы полей Товар.Код и Клиент.Код – строковые.

8. Дан файл sales.txt. Требуется получить результирующий набор данных следующей структуры.

Таблица 4 – Результирующий набор данных

Клиент.Код	Группа 1	Группа 2	Группа 3
0000009	40	39	47
0000049	32	22	31
0000051	545	798	340
...	...	...	...

В ячейках на пересечении групп и клиентов находится количество товара, т.е. например число 39 в таблице означает, что клиент с кодом 0000009 приобрел 39 шт. товаров из группы 2.

Решите задачу двумя способами:

1 С использованием обработчика Слияние.

2 С использованием обработчика Кросс-таблица.

Результаты, выполненные различными способами, должны совпадать.